◎ 1조 이채영 전혜림 최보금 황예진

Stacking ensemble을 이용한 시계열 자료분석

날씨 데이터를 기반으로 한 에너지 소비량 예측 모델링

목차

목차

01 서론

- 이전 연구 고찰
- 문제 제기 & 연구의 동기

1-1. 이전 연구 고찰

논문1

Stacking Ensemble Learning for Short-Term Electricity Consumption Forecasting

- Authors: Federico Divina, Aude Gilson, Francisco Goméz-Vela, Miguel García Torres and José F. Torres
- Received: 2 February 2018; Accepted: 9 April 2018; Published: 16 April 2018
- Energies

논문2

날씨를 고려한 딥러닝 기반의 개별 가구 에너지 사용 요금 예측

- 저자: 박지수, 홍승우, 서일홍
- 2020.4
- 한양대학교 공학 석사학위논문

1-1. 이전 연구 고찰

논문1

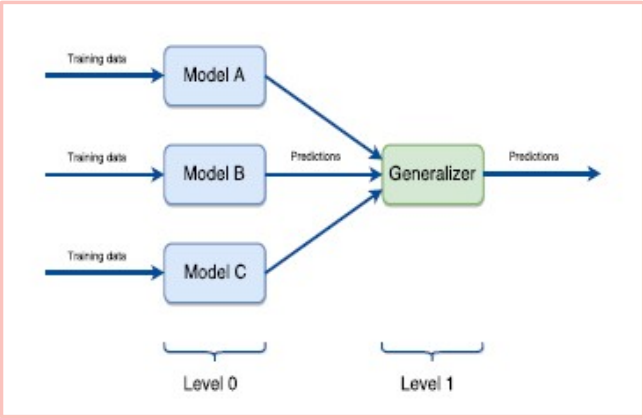
- 최근 분류나 회귀 문제에 딥러닝 기법들 중 개별 모델들의 결과를 합치는 앙상블 기법이 인기가 있다.
- 앙상블은 데이터를 다양한 모델로 훈련시켜 나온 여러 개의 결과값을 하나의 예측값으로 통합하는 기법이다.
- 다양한 앙상블 기법 중 스태킹 기법으로 시계열 데이터에서 에너지 소비량을 예측한다.
- 1단계에서 train data로 부터 여러 개의 base learner가 나오고 그 결과값을 2단계에서 합쳐 하나의 예측값을 도출한다.

논문2

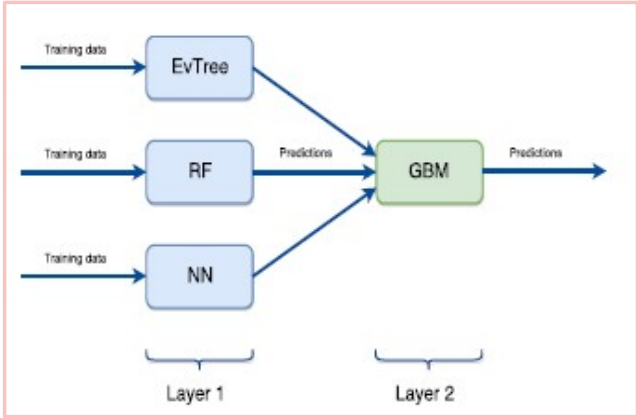
- 에너지 사용 요금은 가계의 고정적인 지출 항목 중 하나로써, 특히 날씨로 인해 에너지 사용이 급증하는 시기에는 높은 누진율이 적용되어 가계 부담을 키우고 있다.
- 합리적인 에너지 사용을 위해 소비자가 고지될 에너지 사용 요금을 사전에 예측하고, 그에 따라 에너지 사용 조절 필요성이 대두된다.
- 딥러닝 기반의 모델을 이용하여 에너지 요금 예측에 큰 영향을 미치는 날씨를 고려한 개별 가구의 월 에너지 사용 요금 예측 방법을 제안한다.

1-1. 이전 연구 고찰

논문1: Ensemble learning



general 한 stacking ensemble 방법의 구조



논문에서 제안하는 모델 구조

하나의 모델을 각각 돌렸을 때의 결과보다 더 나은 결과를 얻기 위해 여러 모델을 결합한다.

1-1. 이전 연구 고찰

논문2: 에너지 소비량에 날씨가 미치는 영향

d	MAE(₩)			
	날씨 o	날씨 x	차이	개선율
5	8460	8950	490	5.5%
10	6240	6502	262	4.0%
15	4805	4966	161	3.2%
20	3596	3686	90	2.5%
25	2451	2503	52	2.1%
평균	5110	5321	211	4.0%

날씨 정보 유무에 따른 에너지 사용 요금 MAE

- 개별가구 에너지 사용 요금 예측에 있어, 날씨 정보를 고려하여 예측한 모델이 그렇지 않은 모델보다 평균 4.0% 적은 오차를 보이는 것을 확인할 수 있었다.
- 따라서, 날씨 정보가 에너지 사용 요금 예측에 필요한 정보를 포함한다고 볼 수 있다.

1-2. 문제 제기 & 연구 동기

- 대부분의 논문에서는 에너지 소비량을 예측할 때 '에너지 소비량' 자체만을 가지고 예측하고자 하는 시점 이전의 몇 개의 값을 x 변수로, 현재 시점의 사용량을 y로 사용하여 예측한다.

- 하지만 대부분의 자연 ·사회 현상은 다양한 변수의 영향 아래에 발생한다.
- 따라서 에너지 소비에 영향을 미치는 변수를 고려한 에너지 소비량 예측을 진행하고자 한다.

- 앞선 두번째 논문을 참고하여, 날씨가 에너지 소비량에 영향을 미칠 것이라 예상하고, 날씨 정보를 활용한 에너지 소비량 예측을 시도한다.

목차

02 본론

- 데이터 소개
- EDA
- 데이터 전처리
- 사용한 방법론 소개

2-1. 데이터 소개

- 3년 3개월 간의 런던 에너지 소비량(1일 단위) & 날씨 데이터
- Dataset : Smart meter data from London area (Smart meters in London)
- 출처 : Kaggle 사이트, https://www.kaggle.com/jeanmidev/smart-meters-in-london

energyuse	에너지 사용량
temperatureMax	전체 최고 기온
temperatureMin	전체 최저 기온
apparentTemperatureMax	최고 체감 온도
apparentTemperatureMin	최저 체감 온도
windBearing	풍력
dewpoint	이슬점
pressure	기압
visibility	가시성

humidity	습도
unIndex	자외선 지수
moonPhase	달의 위상
precipType	rain/snow
icon	날씨정보
Temperature_MaxMin_diff	최고 기온과 최저 기온 시간 차이의 절대값
sunset_sunrise_diff	일몰과 일출 시간 차이의 절대값

2-2. EDA

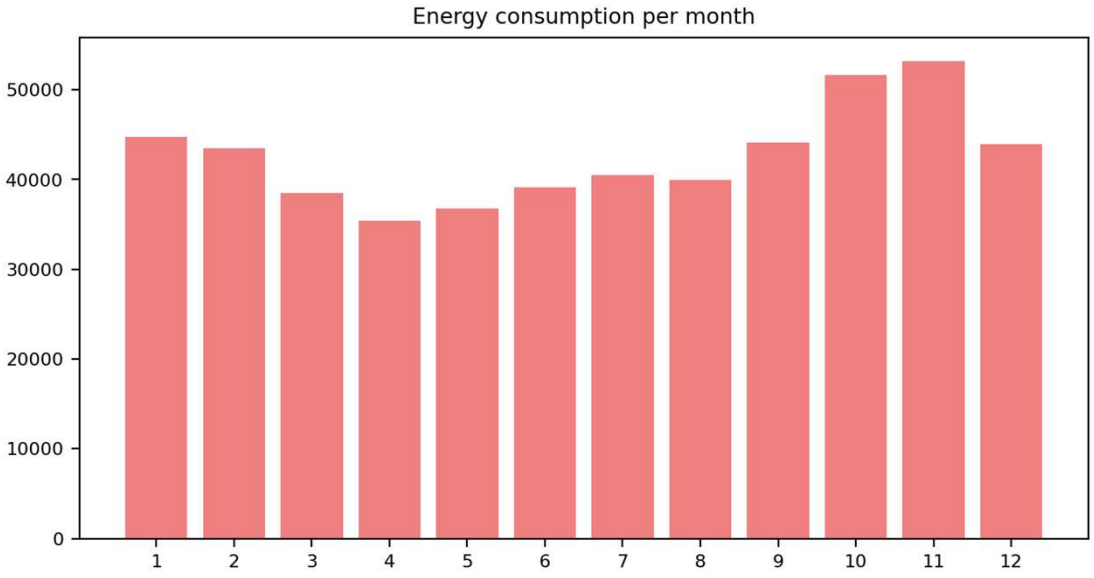
시간에 따른 세 종류의 기온 분포

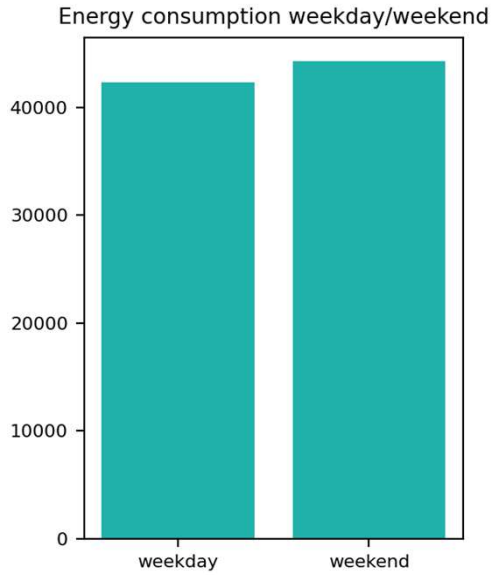
- 기온의 종류는 최고, 낮 기간 최고, 체감 최고 온도로 총 3가지이다. (최저도 마찬가지)
- 왼쪽 두 그래프와 같이 기온의 3가지 종류를 겹쳐서 그려봤을 때 거의 비슷한 양상을 보이기 때문에 실제 최고(저) 기온을 제외한 2가지 기온 변수는 제거한다.

2-2. EDA

월별 평균 에너지 소비량과 주말여부별 평균 에너지 소비량

- 10,11월의 평균 에너지 소비량이 상대적으로 높고 3,4,5월이 낮다.
- 주말보다는 주중 평균 에너지 소비량이 상대적으로 낮지만 큰 차이가 없다.

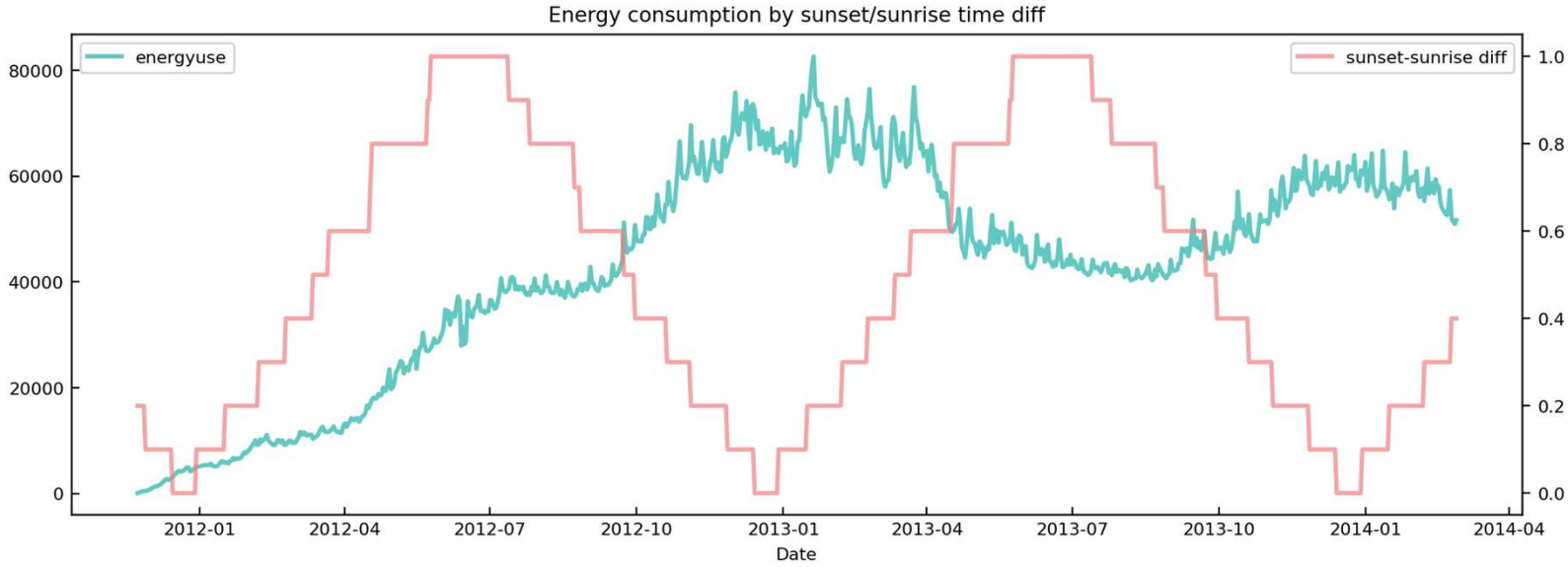




2-2. EDA

일출-일몰 시차와 에너지 소비량 비교

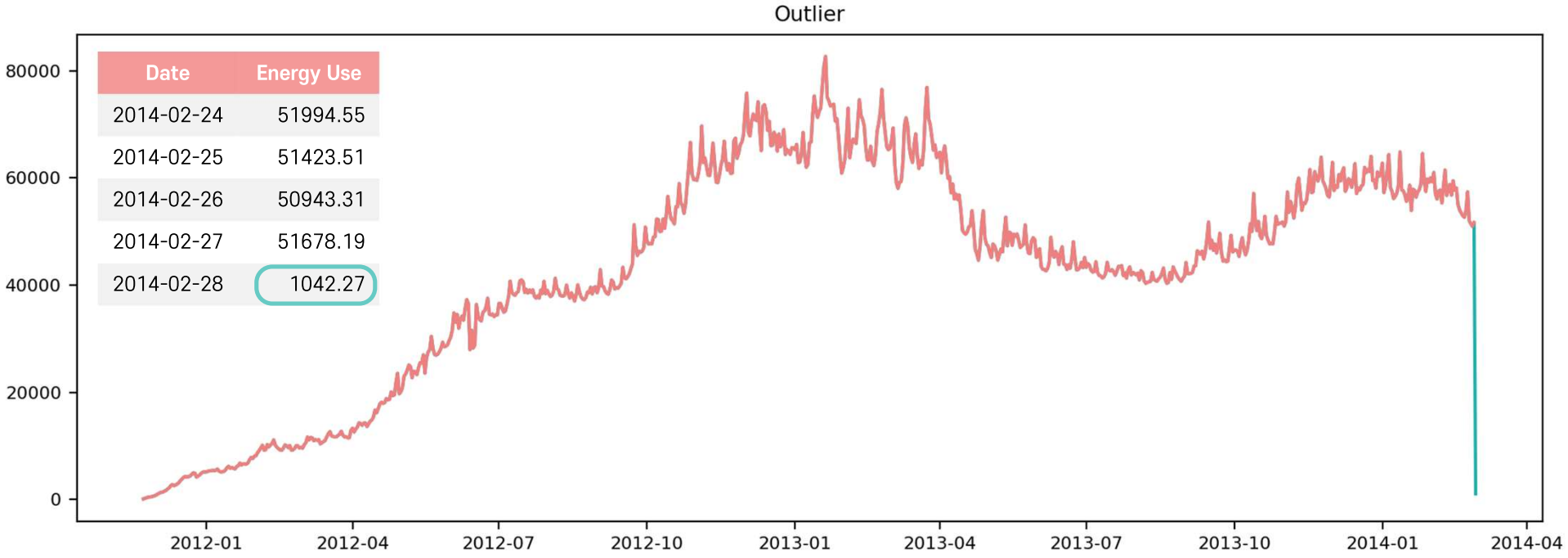
- 일출-일몰 시차는 년도 별로 비슷한 양상을 보인다.
- 에너지 소비량도 증감하는 규칙성을 보이나, 같은 월을 비교했을 때 해를 거듭할수록 에너지 소비량이 증가함을 알 수 있다.
- 시차가 가장 큰 7월에 에너지 소비량이 작고 시차가 가장 작은 1월에 에너지 소비량이 크다(반비례).



2-3. 데이터 전처리

- 2014년 2월 28일의 에너지 소비량의 값이 이전 날들의 값과 크게 차이가 나므로 이상치로 간주하여 제거한다.

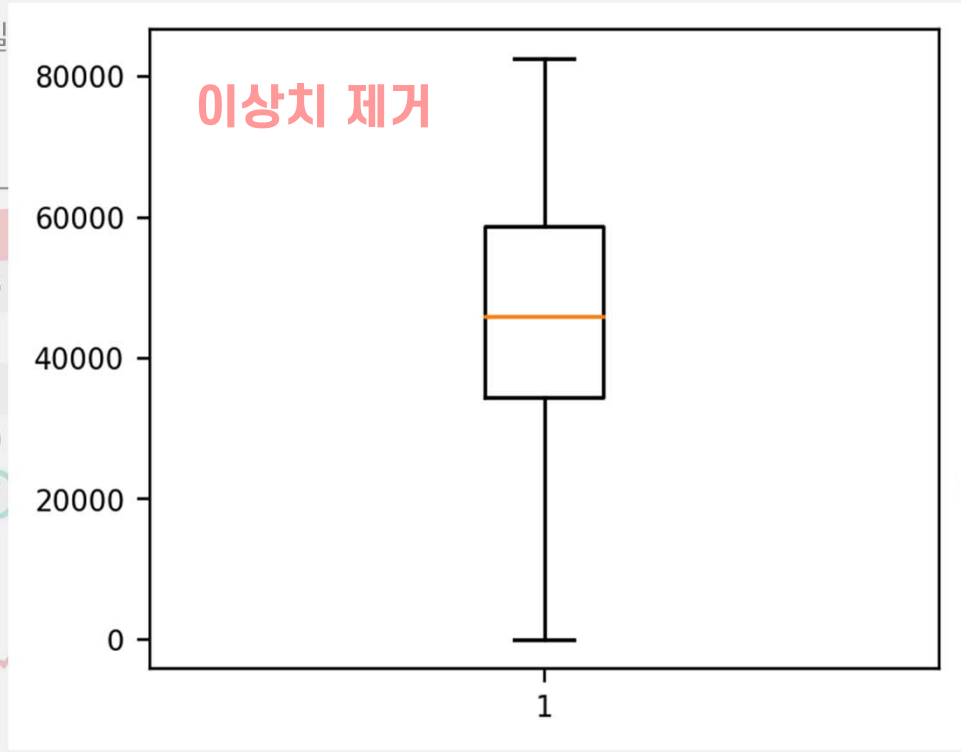
Date	Energy Use
2014-02-24	51994.55
2014-02-25	51423.51
2014-02-26	50943.31
2014-02-27	51678.19
2014-02-28	1042.27



2-3. 데이터 전처리

- 2014년 2월 28일 치로 간주하여 제거한다.

Date	Energy Use
2014-02-24	51994.55
2014-02-25	51423.51
2014-02-26	50943.31
2014-02-27	51678.19
2014-02-28	1042.27



2-3. 데이터 전처리

- Block별로 나눠져 있던 에너지 데이터를 일별 소비량 총합 데이터 하나로 통합

block_0	
day	energy-sum
2012-10-12	7.098
2012-10-13	11.087
...	...

block_1	
day	energy-sum
2012-03-06	11.354
2012-03-07	18.531
...	...

...

block_111	
day	energy-sum
2011-12-07	11.658
2011-12-08	21.522
...	...

energyuse	
day	energy-sum
2011-11-23	90.385
2011-11-24	213.412
...	...

- 분석에 사용하지 않는 column 삭제 (Date, temperatureHigh, temperatureLow 등)

2-3. 데이터 전처리

- 최고, 최저온도 시간 차이와 일몰, 일출 시간 차이의 절대값 변수 생성

weather.csv			
Temperature MaxTime	Temperature MinTime	sunsetTime	sunriseTime
14:00:00	7:00:00	16:03:50	7:32:38
12:00:00	2:00:00	16:02:48	7:34:14
...	...	...	...

weather.csv	
temperature MaxMin_diff	sunset_ sunrise_diff
7	9
10	9
...	...

- 연속형 변수에 Min-Max scaling, 범주형 변수에 One-Hot Encoding 적용

weather.csv			
dew Point	wind Speed	icon	precip Type
6.29	2.04	fog	rain
6.96	5.75	wind	rain
...	...	...	...

weather.csv							
dew point	wind Speed	icon_ clear_day	...	icon_ wind	precipType _rain	...	precipType _snow
0.552	0.189	0	...	0	1	...	0
0.640	0.393	0	...	1	1	...	0
...	...	...	...		...	...	...

2-3. 데이터 전처리

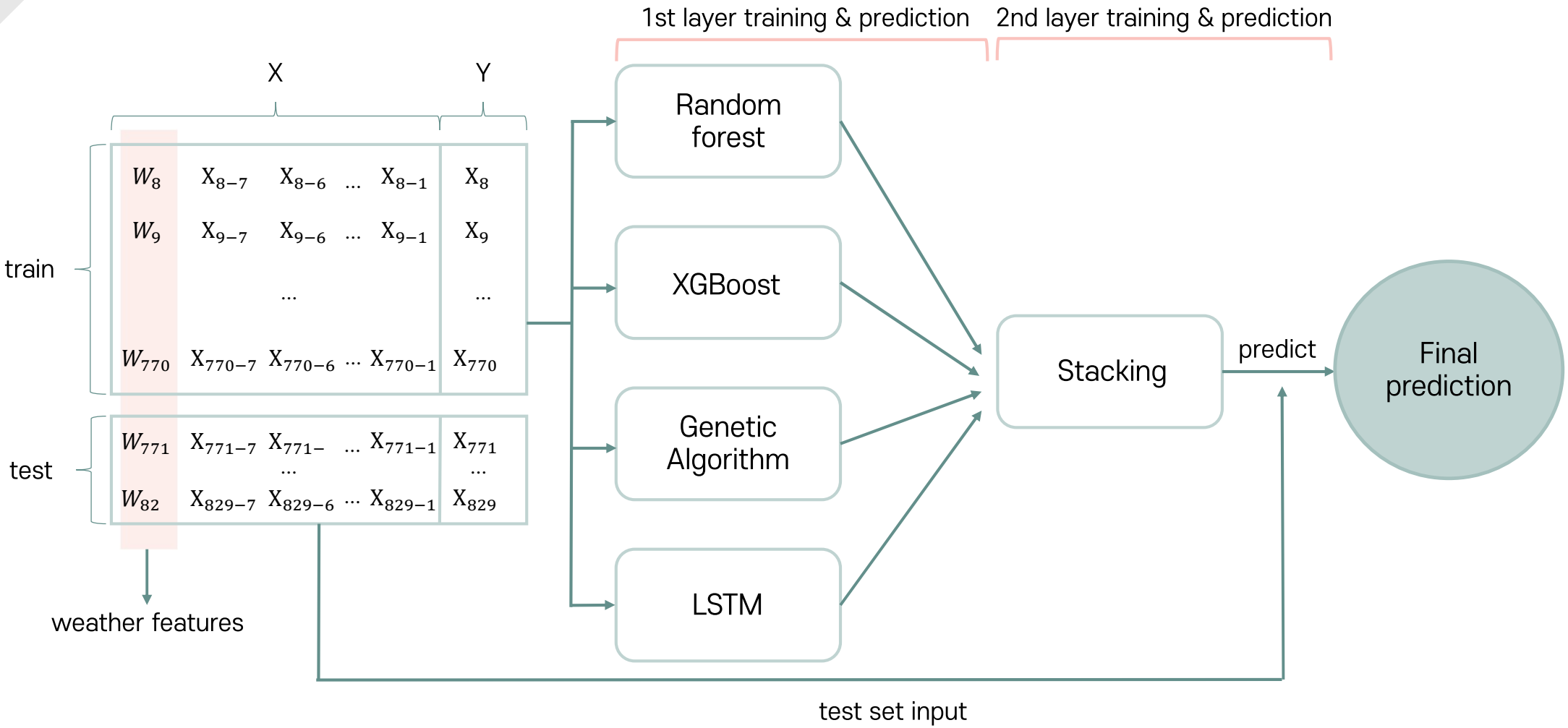
- 해당 시점 7일 전까지의 에너지 사용량 변수를 생성하여 predictor로 사용

energyuse
90.385
213.412
303.993
420.976
444.883
500.686
584.317
669.827
848.949
1014.591
…

shift1	shift2	shift3	shift4	shift5	shift6	shift7
NaN	NaN	NaN	NaN	NaN	NaN	NaN
90.385	NaN	NaN	NaN	NaN	NaN	NaN
213.412	90.385	NaN	NaN	NaN	NaN	NaN
303.993	213.412	90.385	NaN	NaN	NaN	NaN
420.976	303.993	213.412	90.385	NaN	NaN	NaN
444.883	420.976	303.993	213.412	90.385	NaN	NaN
500.686	444.883	420.976	303.993	213.412	90.385	NaN
584.317	500.686	444.883	420.976	303.993	213.412	90.385
669.827	584.317	500.686	444.883	420.976	303.993	213.412
848.949	669.827	584.317	500.686	444.883	420.976	303.993
1014.591	848.949	669.827	584.317	500.686	444.883	420.976

NA값이 없는 7행부터 데이터로 사용

2-4. 사용한 방법론 소개


1st layer training & prediction
2nd layer training & prediction
X
Y
train
test
W_8 X_{8-7} X_{8-6} ... X_{8-1} X_8
W_9 X_{9-7} X_{9-6} ... X_{9-1} X_9
... ...
W_{770} X_{770-7} X_{770-6} ... X_{770-1} X_{770}
W_{771} X_{771-7} X_{771-} ... X_{771-1} X_{771}
W_{82} X_{829-7} X_{829-6} ... X_{829-1} X_{829}
weather features
Random forest
XGBoost
Genetic Algorithm
LSTM
Stacking
predict
Final prediction
test set input

2-4. 사용한 방법론 소개

이전 사용량만을 사용한 예측

X_{8-7}	X_{8-6}	...	X_{8-1}	X_8
X_{9-7}	X_{9-6}	...	X_{9-1}	X_9
	...			...
X_{770-7}	X_{770-6}	...	X_{770-1}	X_{770}
X_{771-7}	X_{771-6}	...	X_{771-1}	X_{771}
	...			...
X_{829-7}	X_{829-6}	...	X_{829-1}	X_{829}

이전 사용량과 날씨 변수를 사용한 예측

W_8	X_{8-7}	X_{8-6}	...	X_{8-1}	X_8
W_9	X_{9-7}	X_{9-6}	...	X_{9-1}	X_9
		...			...
W_{770}	X_{770-7}	X_{770-6}	...	X_{770-1}	X_{770}
W_{771}	X_{771-7}	X_{771-6}	...	X_{771-1}	X_{771}
		...			...
W_{829}	X_{829-7}	X_{829-6}	...	X_{829-1}	X_{829}

날씨 변수를 사용하지 않은 예측 결과와 사용한 예측 결과를 비교해본다.

2-4. 사용한 방법론 소개

1st layer: 개별 모델 학습

Random Forest

데이터를 배깅 방식으로 샘플링하여
결정 트리를 개별 학습한 뒤
최종적으로 보팅을 통해 예측을 결정하는 앙상블 모델.

사용 파라미터: n_estimators = 50

XGBoost

트리 기반의 앙상블 학습 방법으로,
GBM의 느린 수행시간 및 과적합 문제를 해결하는 모델.

사용 파라미터: n_estimators = 1000

Genetic Algorithm

생물학적 진화를 기반으로 하여 적절한 해를 찾아 나가는 모델.

사용 파라미터: function_set=['add', 'sub', 'mul', 'div', 'sqrt', 'log', 'abs', 'neg', 'inv', 'max', 'min', 'sin', 'cos', 'tan'], generations=30

LSTM

RNN의 장기 의존성 문제를 해결하는 순환 신경망의 한 방법.

사용 layer: LSTM(128) -> Dropout(0.2) -> LSTM(128) -> Dropout(0.2) -> Dense

2-4. 사용한 방법론 소개

2nd layer: Stacking

Stacking이란?

개별적인 여러 알고리즘을 서로 결합해 예측 결과를 도출하는 앙상블 방법.

개별 기반 모델의 예측 데이터를 학습 데이터로 만들고, 이를 이용하여 다시 예측을 수행한다.

Method 1	Method 2
개별 기반 모델의 예측 결과값에 평균을 취하여 최종 예측 수행	개별 기반 모델의 예측 결과값에 ML 알고리즘을 수행하여 최종 예측 수행 (AdaBoost, Gradient Boosting, ExtraTree)

AdaBoost

분류 기반의 부스팅 앙상블 모델로, 오류 데이터에 가중치를 부여하며 부스팅을 수행

GradientBoost

분류 기반의 부스팅 앙상블 모델로, 가중치 업데이트로 경사 하강법을 이용

ExtraTree

앙상블 모델인 랜덤 포레스트에서 무작위성을 더 추가한 방식

목차

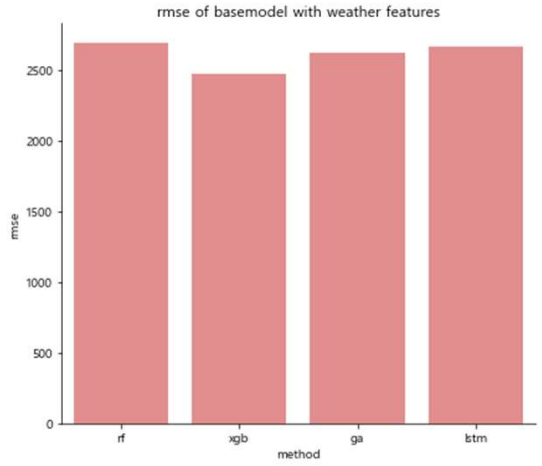
03 결론

- 데이터 분석 결과 제시
- 결과 해석

3-1. 데이터 분석 결과 제시

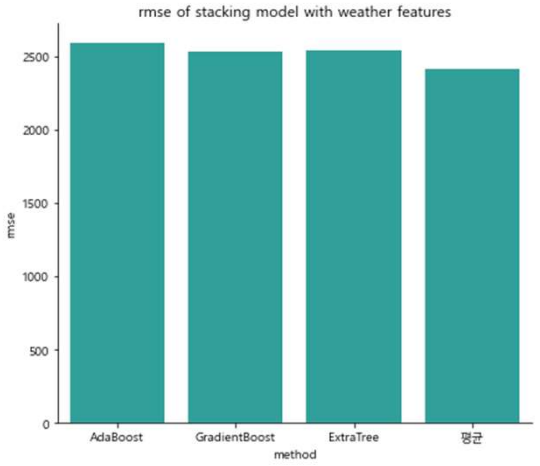
- 날씨 변수까지 포함한 경우

Base model (RMSE)



Random Forest	XGBoost	Genetic Algorithm	LSTM
2692.42	2469.20	2622.96	2666.66

Ensemble model (RMSE)

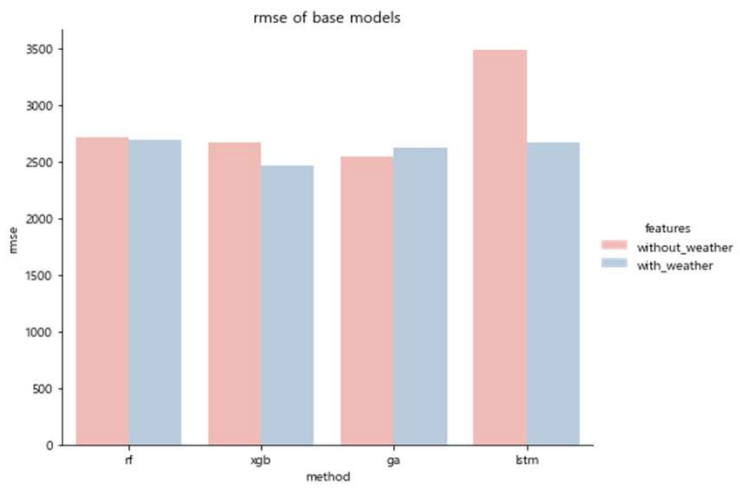


AdaBoost	Gradient Boost	ExtraTree	평균
2594.18	2530.29	2538.89	2413.58

3-1. 데이터 분석 결과 제시

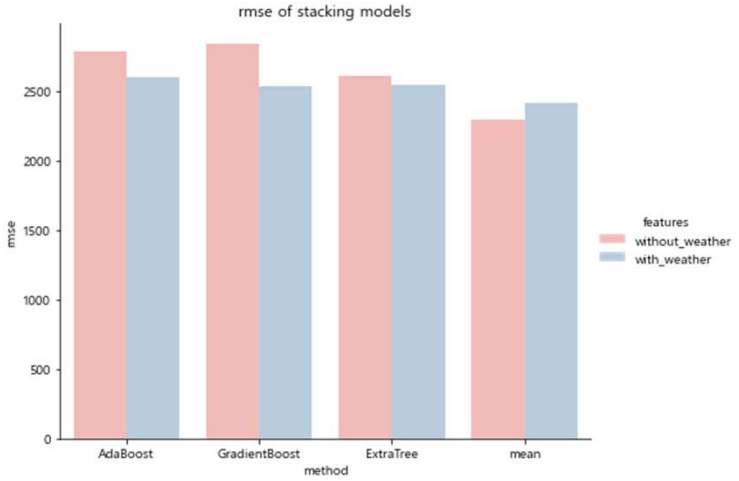
- 날씨 변수를 제외한 경우와 RMSE값 비교

Base model (RMSE)



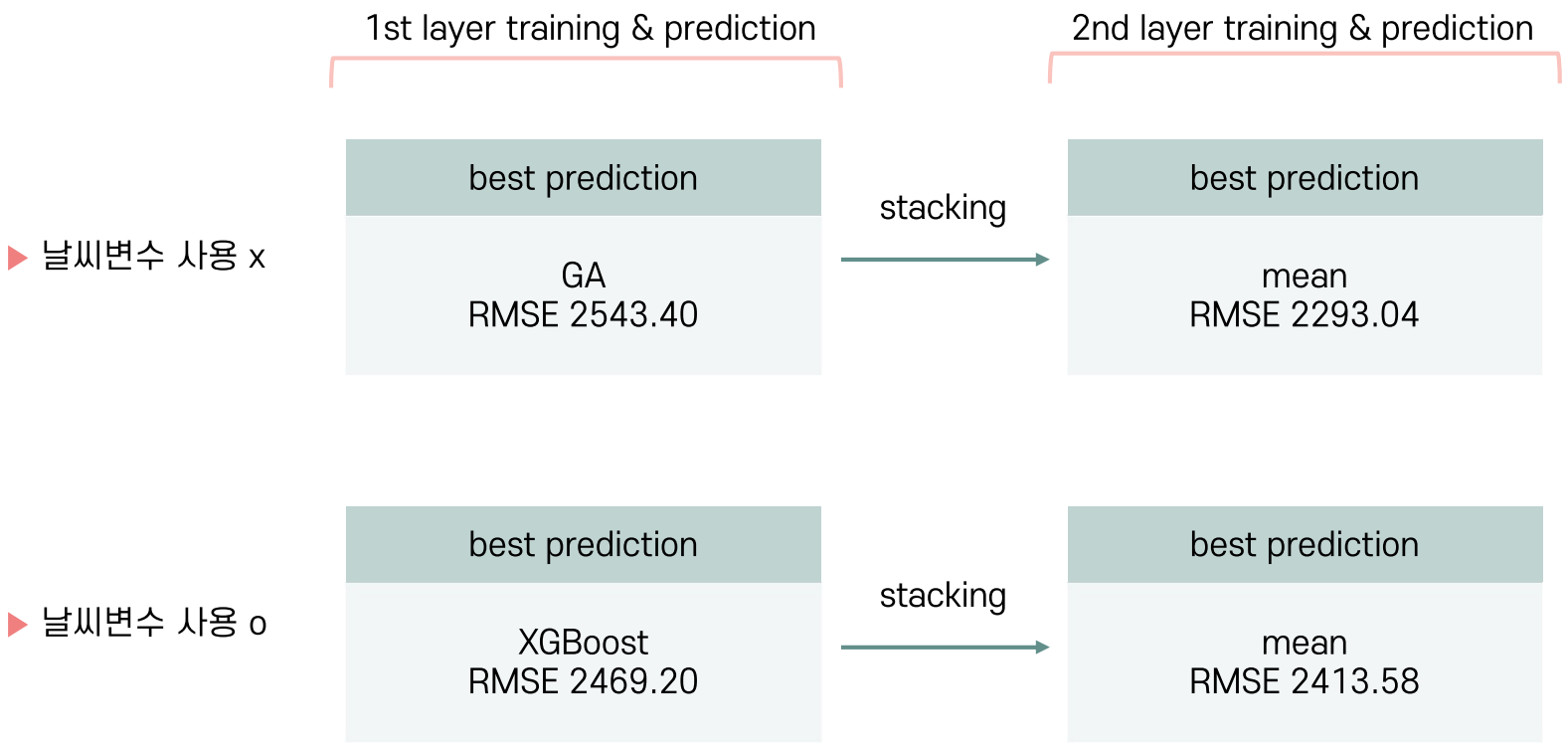
	Random Forest	XGBoost	Genetic Algorithm	LSTM
날씨 x	2714.82	2669.14	2543.40	3492.00
날씨 o	2692.42	2469.20	2622.96	2666.66

Ensemble model (RMSE)



	AdaBoost	Gradient Boost	ExtraTree	평균
날씨 x	2781.63	2839.91	2601.98	2293.04
날씨 o	2594.18	2530.29	2538.89	2413.58

3-1. 데이터 분석 결과 제시



3-2. 결과 해석

Base model (1차 layer)

에너지소비량의 경우 대체로 시간이 지남에 따라 증가하는 추세를 보인다.

➡ 데이터 수집의 초반이었기 때문에 에너지소비량을 측정하는 미터기를 설치한 가정이 적어 초반에는 가정에서 측정된 에너지소비량이 다소 작게 나왔다고 판단했다.

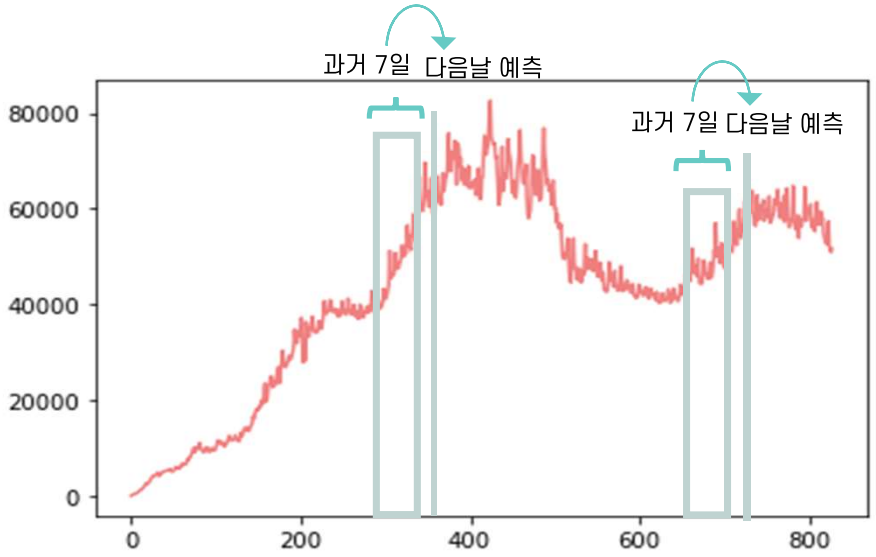
날씨변수를 추가하여 에너지소비량을 예측했을 때, RandomForest와 Xgboost 모델의 RMSE 값이 과거 7일의 에너지소비량만 사용했을 때보다 줄어들었다. 하지만 감소폭이 생각했던 것보다 작았다.

➡ 모델링에 고려했던 날씨변수들이 생각보다 에너지소비량에 유의미한 영향을 주지 못했다. 이는 날씨변수들간 상관성이 높아 발생한 결과라고 판단했다.

3-2. 결과 해석

Stacking ensemble model (2차 layer)

Stacking의 2차 layer에 별도의 ensemble 모델이 아닌 1차 base model들의 예측값의 평균으로 최종 결과값을 도출한 경우가 RSME값이 가장 작게 나왔다.



- 분석에 이용하는 에너지소비량 데이터가 시간에 따라 대체로 증가하는 선형성을 갖는다.
- 따라서 별도의 앙상블 기법을 사용하지 않고 과거 7개의 에너지소비량을 이용해 예측한 경우가 에너지소비량의 최근 변화량을 더 잘 반영하여 RMSE값이 더 작게 나왔다고 판단했다.

목차

04 논의

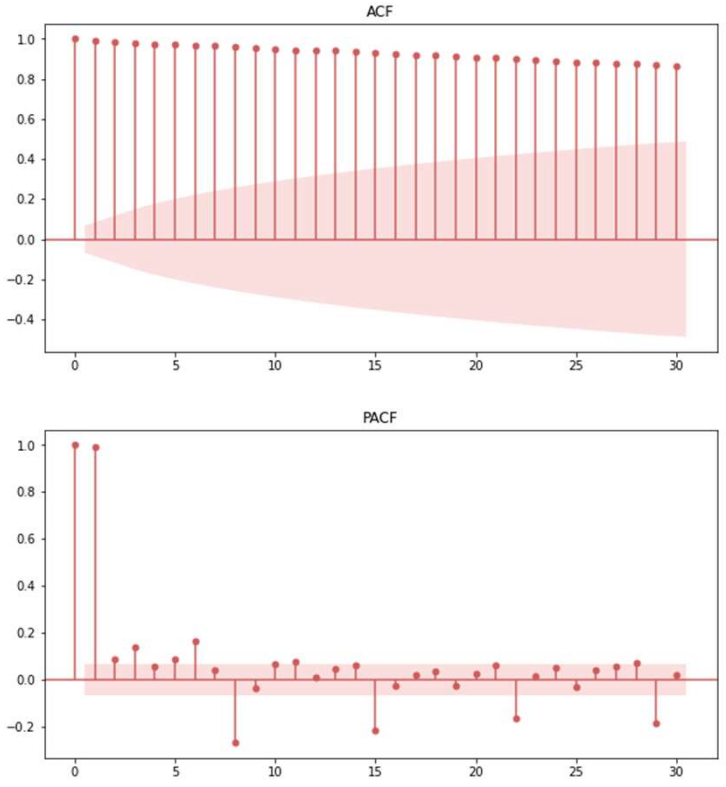
- 분석의 한계
- 추후 연구 방향 제시

4-1. 분석의 한계

- 에너지소비량을 예측하기 위해 다양한 날씨 변수를 추가했지만 유의미한 결과를 얻기 힘들었다.
- 시계열변수들의 추세가 강한 것인지 분석에 고려했던 날씨 변수들이 적합하지 않았던 것인지 판단이 어려워 추가적인 모델링에 어려움을 겪었다.

- ACF의 추세가 느리게 감소하고 PACF의 추세가 경계를 크게 벗어나므로 정상성을 나타내지 않음을 알 수 있다.

비정상(non-stationary) 시계열



4-2. 추후 연구 방향 제시

에너지소비량의 경우 대체로 시간이 지남에 따라 증가하는 추세를 보인다.

➡ 데이터 수집의 초반이었기 때문에 에너지소비량을 측정하는 미터기를 설치한 가정이 적어 초반에는 가정에서 측정된 에너지소비량이 다소 작게 나왔다고 판단했다.

➡ 에너지소비량이 다소 작게 나온 초기 데이터를 제거한 후 동일한 데이터 분석을 하여 나오는 결과를 비교한다.

날씨변수를 추가하여 에너지소비량을 예측했을 때, RandomForest와 Xgboost 모델의 RMSE 값이 과거 7일의 에너지소비량만 사용했을 때보다 줄어들었다. 하지만 감소폭이 생각했던 것보다 작았다.

➡ 모델링에 고려했던 날씨변수들이 생각보다 에너지소비량에 유의미한 영향을 주지 못했다. 이는 날씨변수들간 상관성이 높아 발생한 결과라고 판단했다.

➡ 날씨변수들 간의 상관도를 고려하여 변수 선택을 한 후 모델링하여 나오는 결과를 비교한다.

감사합니다